clustering

clustering:

Clustering is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities.

Advantages:

The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

Disadvantages:

Application:

→ clustering analysis is broadly used in market research, pattern recognition, data analysis and image processing.

→ clustering can also help marketers discover distinct groups in their customer base.

→ In the field of biology, it can be used to derive plant and animal taxonomies.

→ clustering also helps in classifying documents on the web for ~~im~~ information discovery.

→ centroid
medoid

Centroid based clustering: 72 page

In centroid based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set.

k-means clustering is centroid based clustering.

K-means is an iterative clustering algorithm in which items are ~~A~~ moved among sets of clusters until the desired set is reached.

The cluster mean of $K_i = \{t_{i1}, t_{i2}, \cdots t_{im}\}$ is defined as.

$$m_i = \frac{1}{m}\sum_{j=1}^{m} t_{ij}$$

Example 5·4

Suppose that we ^are^ given the following items to cluster:

$\{2, 4, 10, 12, 3, 20, 30, 11, 25\}$

Given, $k = 2$,

$m_1 = 2$

$m_2 = 4$

$\frac{m_1 + m_2}{} \Rightarrow \frac{2+4}{} = \frac{6}{2} = 3$

k_1 is including number ≤ 3

m_1	m_2	k_1	k_2
2	4 [illegible]	$\{2, 3\}$	$\{4, 10, 11, 12, 20, 25, 30\}$
$\frac{3+2}{2}$ 2·5	16 $\frac{\{2_1+10+11..\}}{7}$	$\{2, 3, 4\}$	$\{10, 11, 12, 20, 25, 30\}$
3	18	$\{2, 3, 4, 10\}$	$\{11, 12, 20, 25, 30\}$
4·75	19·6	$\{2, 3, 4, 10, 11, 12\}$	$\{20, 25, 30\}$
7	25	$\{2, 3, 4, 10, 11, 12\}$	$\{20, 25, 30\}$

So, the ~~ta~~ last two steps are identical. $ This will yield identical means, and thus the means have converged.

Our answer is thus

$k_1 = \{2, 3, 4, 10, 11, 12\}$

and $k_2 : \{20, 25, 30\}$

Dendrogram:

(i) Single link:

Smallest distance between an element in one cluster and an element in the other.

We thus have $dis(k_i, k_j) = min(dis(t_{i1}, t_{jm})) \forall t_{ij} \in k_i \notin k_j$ and $\forall t_{jm} \in k_j \notin k_i$.

(ii) Complete link:

largest distance between an element in one cluster and an element in the other.

We thus have $dis(k_i, k_j) = max(dis(t_{i1}, t_{jm})) \forall t_{i1} \in k_i \notin k_j$ and $\forall t_{jm} \in k_j \notin k_i$.

(iii) Average:

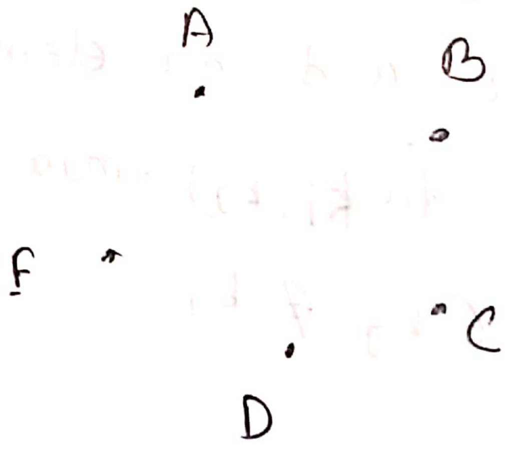
Average distance between an element in one cluster and an element in the other.

we thus have $dis(k_i, k_j) = mean(dis(t_{i1}, t_{jm})) \forall t_{ij} \in$ and $t_{tjm} \in k_j \in k_i$.

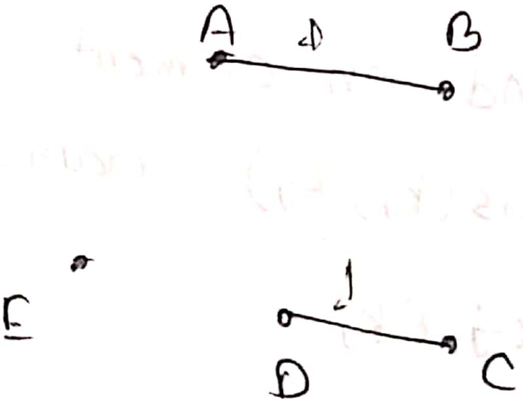
Example 5.3

item	A	B	c	D	E
A	0	1	2	2	3
B	1	0	2	4	3
c	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

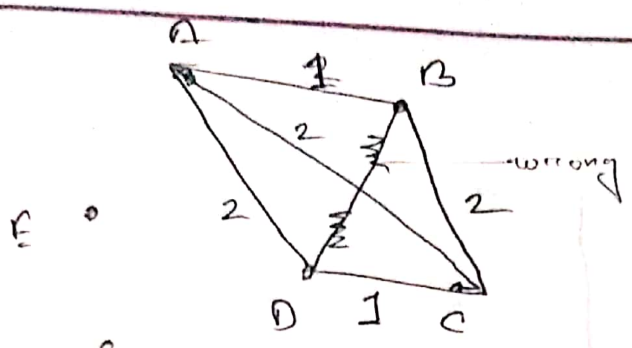
Single Link:



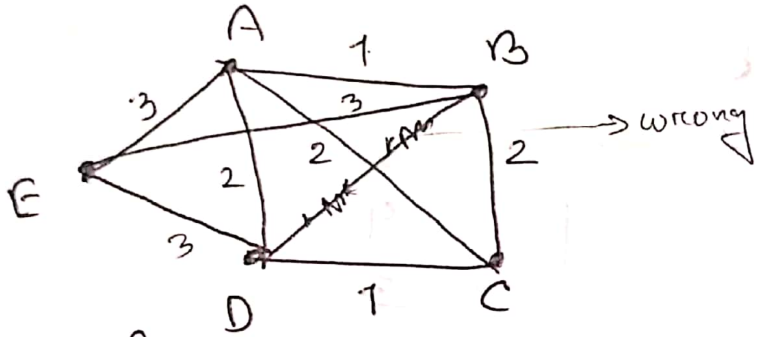
for threshold 0



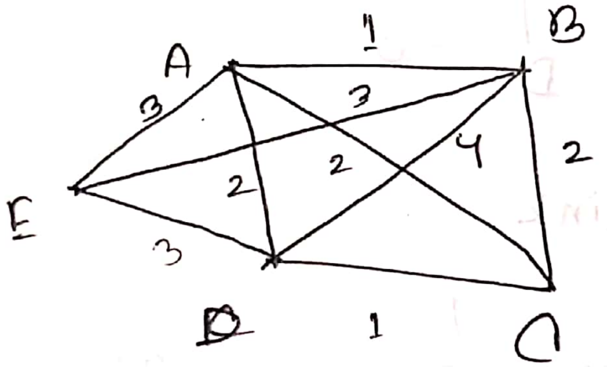
for threshold 1



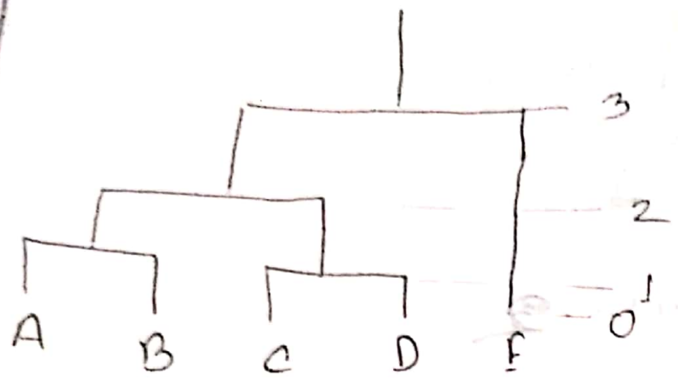
for threshold 2



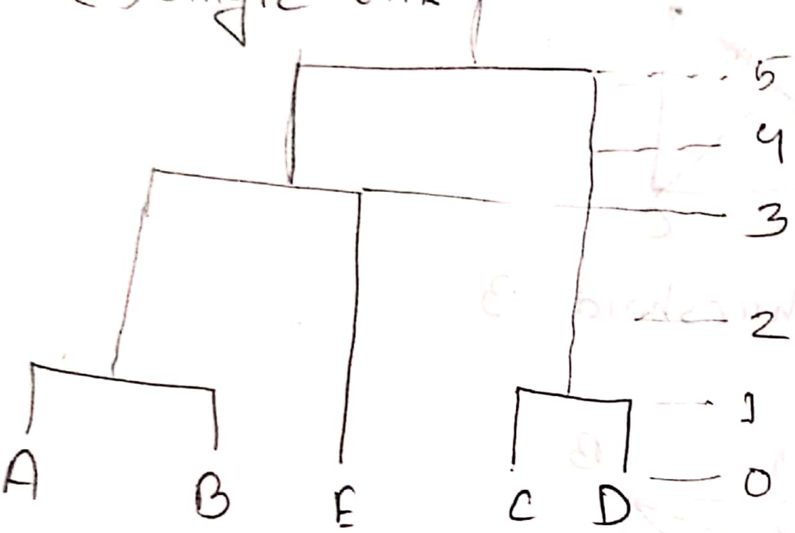
for threshold 3



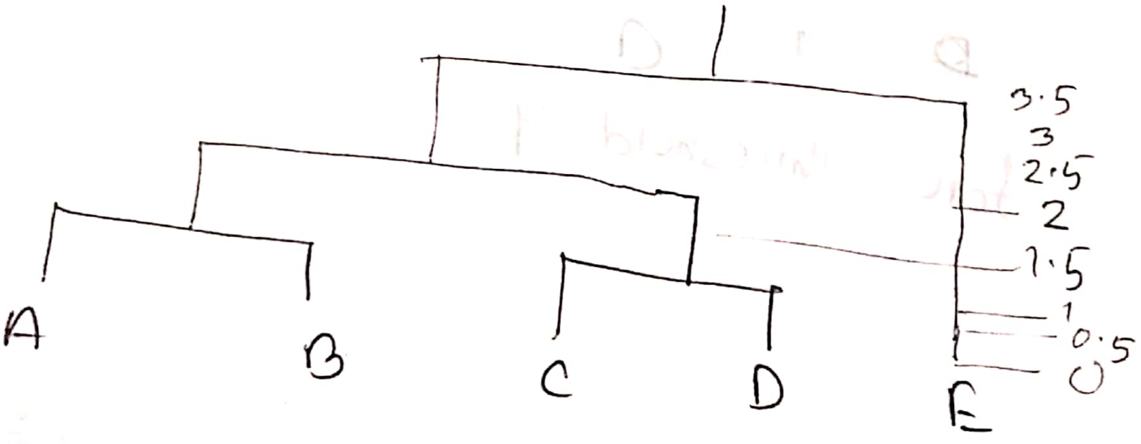
for threshold 4



(a) single link



(b) complete link



(c) Averge link .

k nearest neighbor Algorithm!

$k_1 = \{A\}$

The distance between A to B $= 1 \leq 2$

$\therefore k_1 = \{A, B\}$

The distance between A to C $: 2 \leq 2$

" " " B to C $: 2 \leq 2$

$\therefore k_1 = \{A, B, C\}$

" " " A to D $: 2 \leq 2$

" " " B to D $: 4 \not\leq 2$

" " " C to D $: 1 \leq 2$

" " " $\therefore k_1 = \{A, B, C, D\}$

$\therefore$ " " " A to E $= 3$

B to E $= 3$

C to E $= 5$

D to E $= 3$

$\therefore k_1 = \{A, B, C, D\}$

$\therefore k_2 = \{E\}$

☐ BIRCH:

Balanced iterative reducing and clustering using hierarchies) is designed for clustering a large amount of metric data. It is applies only to numeric data.

□ A clustering feature (CF) is a triple $(N, \overline{LS}, SS)$ where the number of the points in the cluster is N, $\overline{LS}$ is the sum of the points in the cluster, and SS is the sum of the eque squares of the points in the cluster.

□ CF tree:

A CF tree is a balanced tree with a branching factor (maximum number of children a node may have) B.

Each internal node contains a CF triple for each of its children. Each leaf node also represents a cluster and contains a CF entry for each subcluster in it.

A ~~subscriber~~ subcluster in a leaf node must have a diameter no greater than a given threshold value T.

□ DBSCAN!

(density-based spatial clustering of applications with noise} is to create with a minimum size and density.

Density is defined as a minimum number of points within a certain distance of each other.